Краткое руководство по универсальному принятию

Что значит «универсальное принятие»?

Часть программного обеспечения не способна распознать и корректно обработать все доменные имена и адреса электронной почты. Доменные имена могут включать строки доменов верхнего уровня, длина которых превышает длину строк привычных старых доменов. При этом в доменных именах и адресах электронной почты теперь могут использоваться символы из набора Unicode, который гораздо шире традиционного ASCII[1].

Универсальное принятие (UA) – это состояние, когда все допустимые доменные имена и адреса электронной почты правильно и единообразно **принимаются, проверяются, хранятся, обрабатываются и отображаются**.

Группа управления по универсальному принятию (UASG) – это возглавляемая сообществом инициатива по повышению осведомленности, а также выявлению и решению проблем, связанных с универсальным принятием всех доменных имен и адресов электронной почты. Цель этой инициативы – оказание содействия созданию единообразного интерфейса и положительного опыта взаимодействия для интернет-пользователей всего мира. Она осуществляется при поддержке со стороны ICANN (Интернет-корпорации по присвоению имен и номеров) при участии представителей более 200 организаций со всего мира, в том числе Afilias, Apple, CNNIC, GoDaddy, Google, Microsoft и Verisign. Дополнительные сведения о UASG и последних достижениях представлены здесь: **www.uasg.tech**.

ПРИНЯТИЕ

ПРОВЕРКА

ХРАНЕНИЕ

ОБРАБОТКА

ОТОБРАЖЕНИЕ

Универсальное принятие обеспечивается программным обеспечением и онлайн-сервисами, если они поддерживают перечисленные выше возможности для всех доменов и адресов электронной почты.

В **настоящем кратком руководстве** представлены рекомендации UASG по обеспечению универсального принятия с точки зрения пяти аспектов взаимодействия систем с доменными именами и адресами электронной почты: принятие, проверка, хранение, обработка и отображение. Документ предназначен для руководителей высшего звена и менеджеров, отвечающих за деятельность в области информационных технологий и разработки программных продуктов. Рекомендации UASG представлены здесь в общем виде без детализации, необходимой архитекторам программного обеспечения или программистам. Эту более детальную информацию можно получить в документе UASG 007 «Основные сведения об универсальном принятии».

[1]ASCII – традиционно используемая в интернете система кодирования символов, которая определена интернет-стандартом RFC 20 (https://tools.ietf.org/html/rfc20). Набор Unicode определен Консорциумом Unicode (http://unicode.org).

ПРИНЯТИЕ

Принятие – это процесс получения доменного имени или адреса электронной почты от того или иного пользовательского интерфейса, файла или API, используемого приложением или интернет-сервисом.

Рекомендации UASG

- Поля ввода должны быть достаточно большими, чтобы принять все допустимые входные данные. В зависимости от алгоритма кодирования, ввод доменного имени может потребовать до 670 байт. В дополнение к доменному имени адрес электронной почты может содержать локальную часть (перед знаком @) длиной до 64 байт. В результате, общая длина может достигать 735 байт.
- Приложения и сервисы должны распознавать доменные имена и адреса электронной почты в кодировке UTF-8[2] и допускать возможность того, что количество байт кода UTF-8 может превышать количество отображаемых символов.
- IDN-домен может вводиться и отображаться либо с использованием оригинального набора символов, либо с использованием варианта ASCII, поддерживающего обратную совместимость, например, 测试 и xn--0zwm56d. При использовании Unicode для кодировки оригинального набора символов создается U-метка; совместимый с ASCII эквивалент называется A-меткой.[3] Программное обеспечение должно принимать как A-метки, так и U-метки, но при этом оно должно выполнять преобразование A-меток в U-метки при отображении и любой обработке, не требующей применения A-меток.
- Практически во всех случаях введенное доменное имя или адрес электронной почты перед дальнейшей обработкой следует преобразовать в Форму нормализации Unicode C (NFC)[4]. Поскольку при создании NFC происходит некоторая потеря данных, в редких случаях может возникнуть необходимость отложить нормализацию до момента, когда при последующей обработке будет установлен конкретный контекст применения этого алгоритма.

[2] При использовании UTF-8 каждая кодовая точка Unicode кодируется в виде последовательности от одного до четырех байт. Это определено в RFC 3629.

[3] Для прямого и обратного преобразования U-меток в A-метки используется алгоритм Punycode, определенный стандартами RFC 3492 и RFC 5891.

[4] См. стандарт Unicode, приложение № 15 «Формы нормализации Unicode» (https://www.unicode.org/reports/tr15/tr15-47.html).

ПРОВЕРКА

Проверка – это процесс синтаксической проверки адреса электронной почты или доменного имени и, когда это целесообразно, проверки существования доменного имени в DNS. Для работы с современными доменными именами и адресами электронной почты может потребоваться обновление методов проверки.

[5]Определение IDNA2008 см. в стандартах RFC 5890, 5891, 5892, 5893 и 5894.

[6]См. список доменов верхнего уровня (https://www.icann.org/resources/pages/tlds-2012-02-25-en).

Рекомендации UASG

- Введенные данные следует проверять сообразно их целевому назначению. Все доменные имена следует проверять на соответствие стандарту «Интернационализированные доменные имена в приложениях». В настоящее время это стандарт IDNA2008.[5] Проверка проводится для подтверждения синтаксической правильности имени.
- Если ожидается, что вводимая строка – существующая в DNS запись, ее необходимо проверить путем поиска по DNS.
- Если ожидается, что вводимая строка – это допустимое доменное имя, которое (пока) отсутствует в DNS, часть его все равно можно проверить. Например, доменное имя верхнего уровня (TLD) можно проверить на предмет наличия в официальном списке допустимых TLD, который поддерживается Администрацией адресного пространства интернета (IANA).[6]
- При проверке адреса электронной почты его доменная часть проверяется согласно описанию выше. Поскольку локальная часть адреса электронной почты определяется только принимающей почтовой системой, эту часть обычно нельзя проверить. Просьбы ввести адрес электронной почты дважды могут привести к возникновению опечаток.
- В большинстве случаев для всех компонентов доменного имени или адреса электронной почты (кроме TLD, если это не IDN-домен) следует использовать один алфавит (напр., арабский или ханьский) или тесно связанные алфавиты (напр., японские кандзи, катакана, хирагана и ромадзи). Для проверки того, что алфавиты в последовательности Unicode соответствуют передовой практике, применяется технический стандарт Unicode № 39 «Механизмы безопасности Unicode» (https://unicode.org/reports/tr39/#Restriction_Level_Detection).

ХРАНЕНИЕ

Хранение – временное или длительное хранение доменных имен и адресов электронной почты в четко определенных форматах, независимо от ожидаемого срока хранения.

Рекомендации UASG

- Практически во всех случаях доменные имена и адреса электронной почты перед хранением следует преобразовать в Форму нормализации Unicode C (NFC). Поскольку при создании NFC происходит некоторая потеря данных, в редких случаях может возникнуть необходимость отложить нормализацию до момента, когда при последующей обработке будет установлен конкретный контекст применения этого алгоритма.
- В большинстве приложений доменные имена и адреса электронной почты следует хранить в файлах и базах данных с кодировкой UTF-8. Это наиболее распространенный и лучше всего поддерживаемый формат кодирования Unicode. В некоторых случаях, когда нужно обеспечить функциональную совместимость программного обеспечения со старыми базами данных, более удобным решением может оказаться использование такого же формата кодирования, как и у базы данных.
- Наиболее целесообразное представление Unicode в коде приложения зависит от программного окружения. Многие распространенные языки программирования, в том числе языки сценариев Python и Perl, имеют встроенную поддержку Unicode и автоматического прямого и обратного преобразования данных в UTF-8 при вводе и выводе.
- Следует выбрать единое внутреннее представление IDN-доменов в приложениях – в виде U-меток или A-меток. Поскольку каждую допустимую U-метку можно преобразовать в уникальную A-метку и наоборот, приемлемыми являются оба варианта.

ОБРАБОТКА

Обработка происходит всякий раз, когда адрес электронной почты или доменное имя используется приложением или сервисом для выполнения того или иного действия (например, поиска или сортировки списка) или преобразования в альтернативный формат (например, из старой кодировки в UTF-8). Во время обработки может осуществляться дополнительная проверка.

Рекомендации UASG

- По мере развития Unicode обновляйте программное обеспечение, когда это уместно, чтобы использовать последнюю версию стандарта и все доступные графические элементы и шрифты. Учитывайте, что последняя версия может не поддерживаться пользовательскими устройствами, библиотеками программного обеспечения и веб-стандартами, в связи с чем новые символы могут отображаться или неправильно, или как универсальная рамка (□), или не отображаться совсем.
- При наличии API, поддерживающих ввод и вывод в формате UTF-8, используйте их вместо API, не поддерживающих этот формат. Используйте для обработки и проверки IDN-доменов стандартные хорошо отлаженные библиотеки, такие как GNU libidn2 (https://www.gnu.org/software/libidn/#libidn2); не внедряйте собственные решения.
- Символы языков с письмом справа налево требуют особого внимания при использовании в доменных именах и адресах электронной почты. Некоторые из вышеописанных соображений рассматриваются в документе IDNA[7] (для доменных имен) и приложении к стандарту Unicode[8] (для адресов электронной почты).
- При создании реестров или других структур данных, содержащих информацию об алфавите или языке, обеспечьте максимально возможную, а лучше всего – полную поддержку стандарта Unicode.[9] Помните о том, что в некоторых языках может использоваться несколько наборов символов, и что некоторые наборы символов могут использоваться многими языками.

[7]См. RFC 5893, «Системы письма справа налево в интернационализированных доменных именах для приложений (IDNA)» (https://tools.ietf.org/html/rfc5893).

[8]См. UAX № 9, «Двунаправленный алгоритм Unicode» (http://unicode.org/reports/tr9).

[9]См. «Поддерживаемые наборы символов» Unicode (http://unicode.org/standard/supported.html).

ОТОБРАЖЕНИЕ

Отображение происходит при визуализации адреса электронной почты или доменного имени в пользовательском интерфейсе. При отображении доменных имен и адресов электронной почты обычно не возникает затруднений, если используемые алфавиты и все необходимые механизмы визуализации поддерживаются операционной системой, а формат хранения строк соответствует стандарту Unicode. В противном случае могут потребоваться операции преобразования для конкретного приложения.

Рекомендации UASG

- Следует учесть, что, хотя современные программы и устройства способны отображать практически все кодовые точки Unicode, более старые системы могут обеспечивать ограниченную поддержку и требовать, чтобы приложения обрабатывали их старые шрифты. Кроме того, при добавлении в Unicode новых кодовых точек устройства и приложения не будут их отображать до обновления библиотек шрифтов.
- Используйте для отображения IDN-доменов их родные символы, если нет требования отображать их в виде A-меток.
- Доменные имена и адреса электронной почты могут отображаться в виде текста, написанного слева направо (LTR), как в английском или русском языках, или справа налево (RTL), как в арабском или иврите. Поскольку в Unicode атрибуты направления присваиваются индивидуальным кодовым точкам, а не последовательности кодовых точек, некоторые типы смешанного LTR и RTL («двунаправленного») текста понятны пользователям, а некоторые нет. Используйте критерии уровней ограничения Unicode[10] для выявления потенциально могущих создать путаницу строк.
- Интернет-пользователи читают и говорят на многих языках. В некоторых случаях необходима разработка приложений специально для тех или иных языков или групп языков.

[10]Для проверки того, что алфавиты в последовательности Unicode соответствуют передовой практике, см. технический стандарт Unicode № 39, «Механизмы безопасности Unicode» (https://www.unicode.org/reports/tr39/#Restriction_Level_Detection), где описаны умеренный и высокий уровни ограничения.

Подготовьтесь к универсальному принятию

Проверка исходного кода и тестирование устройств

Программы и системы, разработанные или обновленные для поддержки универсального принятия, следует проанализировать и протестировать, чтобы убедиться в правильности их функционирования и чтобы найти и устранить ошибки. В рамках работы по повышению осведомленности группа UASG устанавливает контакты с разработчиками приложений и поставщиками интернет-сервисов, чтобы привлечь их к анализу и тестированию исходного кода универсального принятия, и распространяет перечень критериев, которые могут использоваться для разработки стандартных сценариев тестирования.

Тестирование

UASG также занимается составлением списка сайтов, приложений, адресов электронной почты и доменных имен, подходящих для тестирования. В некоторых случаях тестирование можно автоматизировать и выполнять без вмешательства человека. Практическим примером является недавнее исследование gTLD, выполненное APNIC Labs по поручению ICANN: https://tinyurl.com/new-gtld-ua. UASG сейчас занимается изучением методов автоматизированного тестирования универсального принятия и сообщит о полученных результатах.

Дополнительная литература

В перечисленных ниже документах содержится дополнительная информация об универсальном принятии, Unicode и интернационализированных доменных именах.

- UASG 007 «Основные сведения об универсальном принятии (https://uasg.tech/documents).
- RFC 5894, «Интернационализированные доменные имена для приложений (IDNA): история вопроса, пояснение и обоснование» (https://www.rfc-editor.org/info/rfc5894).
- «Международная веб-типографика» – графический обзор проблем и проблематики работы с разными языками в интернете (https://w3c.github.io/typography/gap-analysis/language-matrix.html).

О терминологии

Одна из трудностей обеспечения универсального принятия заключается в том, что многие термины и концепции, знакомые людям, привыкшим к простым небольшим наборам отдельных «алфавитных» символов, например к латинскому алфавиту, могут создать большую путаницу, если применять их по отношению к системам письма, основанным на других принципах. Для включения широкого спектра систем письма в область интернационализированных доменных имен (IDN-доменов) потребовалось изобрести новую терминологию и использовать знакомые термины (например, «символ») в узком и очень конкретном смысле. В настоящем кратком руководстве предпринята попытка избежать подобных терминов или предоставлять их определение при использовании, однако изучение других материалов, в том числе упомянутых здесь документов, скорее всего потребует более глубокого понимания терминологии.